Міністерство освіти і науки, молоді та спорту України
Національний технічний університет України
“Київський політехнічний інститут”
Фізико-Технічний інститут

Комп’ютерний практикум №1

Криптографія

Виконали:

Студенти 3-го курсу ФТІ
групи ФБ-93
Тішков М.С. та Папуча Н.В.

Перевірила:

Селюх П.В.

Мета роботи

Засвоєння понять ентропії на символ джерела та його надлишковості, вивчення та порівняння різних моделей джерела відкритого тексту для наближеного визначення ентропії, набуття практичних навичок щодо оцінки ентропії на символ джерела.

Порядок виконання роботи

0. Уважно прочитати методичні вказівки до виконання комп’ютерного практикуму.

1. Написати програми для підрахунку частот букв і частот біграм в тексті, а також підрахунку H_1 та H_2 за безпосереднім означенням. Підрахувати частоти букв та біграм, а також значення H_1 та H_2 на довільно обраному тексті російською мовою достатньої довжини (щонайменше 1Мб), де імовірності замінити відповідними частотами. Також одержати значення H_1 та H_2 на тому ж тексті, в якому вилучено всі пробіли.

2. Використовуючи отримані значення ентропії, оцінити надлишковість російської мови в різних моделях джерела.

1. За допомогою написаною програмою отримаємо значення H1, H2, таблиці з частотами монограмм та біграмм та надишковість.

На вхід подаються два типа обробленого тексту: з пробілами та без

Такі результати були отримані:

Для тексту без пробілів:

Частота монограмм:

	Frequency
о	0,110783
е	0,085526
а	0,083096
т	0,06328
н	0,063122
и	0,062707
с	0,054542
л	0,049586
р	0,045731
в	0,044296
м	0,033788
к	0,033531
д	0,032297
у	0,028324
п	0,02596
я	0,020588
ь	0,020395
ы	0,018903
г	0,018788
б	0,01724
з	0,016917
ч	0,014943
й	0,010579
ж	0,010548
ш	0,007988
х	0,006472
ю	0,005959
э	0,004183
ц	0,003658
ф	0,003161
щ	0,003108

Частота біграмм без перетину (перші 10 найчастіші):

	Frequency
то	0,016521
ст	0,01293
на	0,010936
ал	0,010884
не	0,010644
но	0,010469
по	0,010413
ен	0,010203
ос	0,010126
го	0,009751

Частота біграмм з перетином (перші 10 найчастіші):

	Frequency
то	0,016641
ст	0,012903
на	0,010945
не	0,010761
ал	0,010751
по	0,010457
но	0,010381
ен	0,010228
ос	0,010173
ов	0,009843

Для тексту з пробілами:

Частота монограмм:

	Frequency
	0,858521
о	0,015674
е	0,0121
а	0,011756
т	0,008953
н	0,00893
и	0,008872
с	0,007717
л	0,007015
р	0,00647
в	0,006267
м	0,00478
к	0,004744
д	0,004569
у	0,004007
п	0,003673
я	0,002913
ь	0,002885
ы	0,002674
г	0,002658
б	0,002439
з	0,002393
ч	0,002114
й	0,001497
ж	0,001492
ш	0,00113
х	0,000916
ю	0,000843
э	0,000592
ц	0,000518
ф	0,000447
щ	0,00044

Частота біграмм без перетину (перші 10 найчастіші):

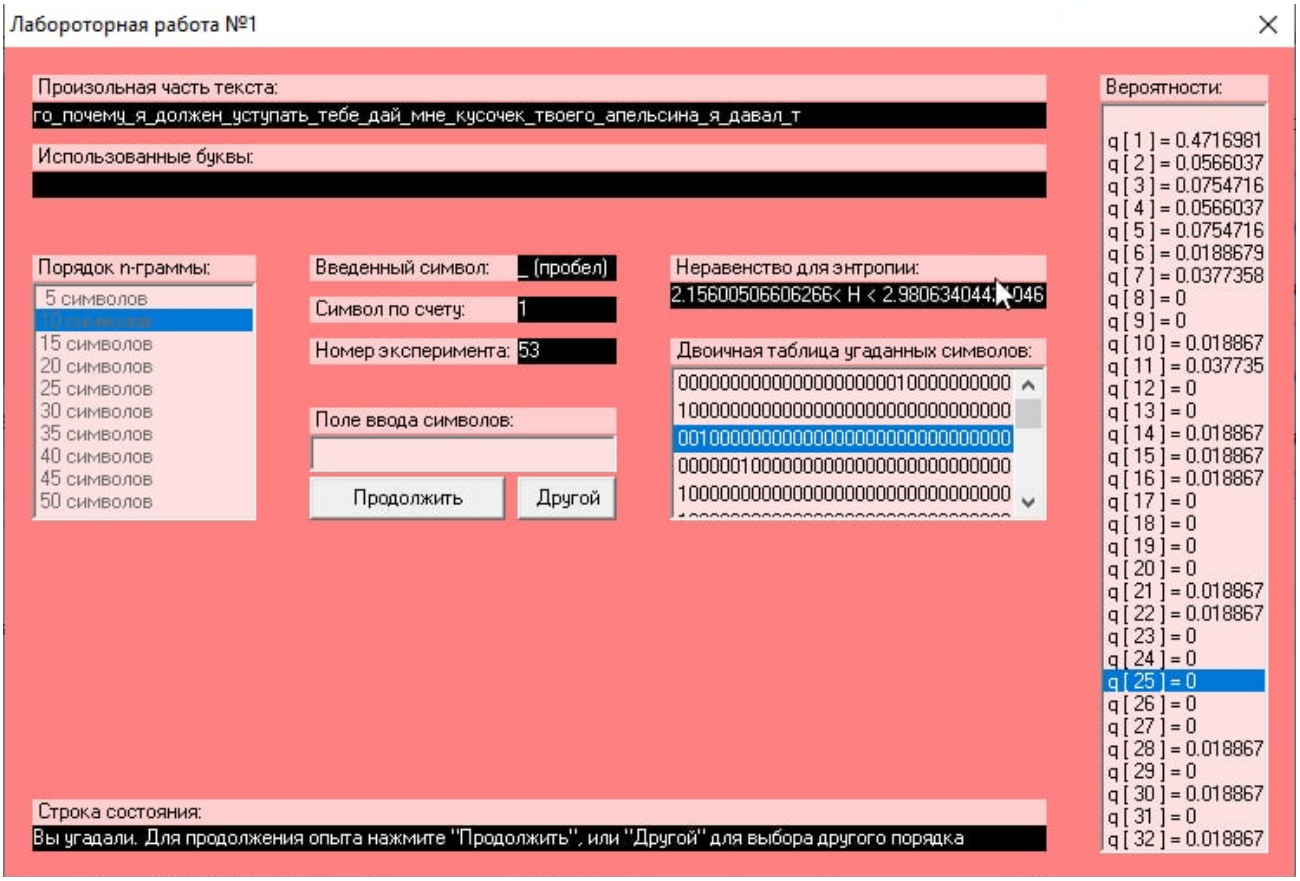
	Frequency
__	0,831722
о_	0,003376
е_	0,002971
_в	0,002837
_с	0,002631
и_	0,00259
а_	0,002537
_п	0,00252
то	0,002313
_н	0,002245

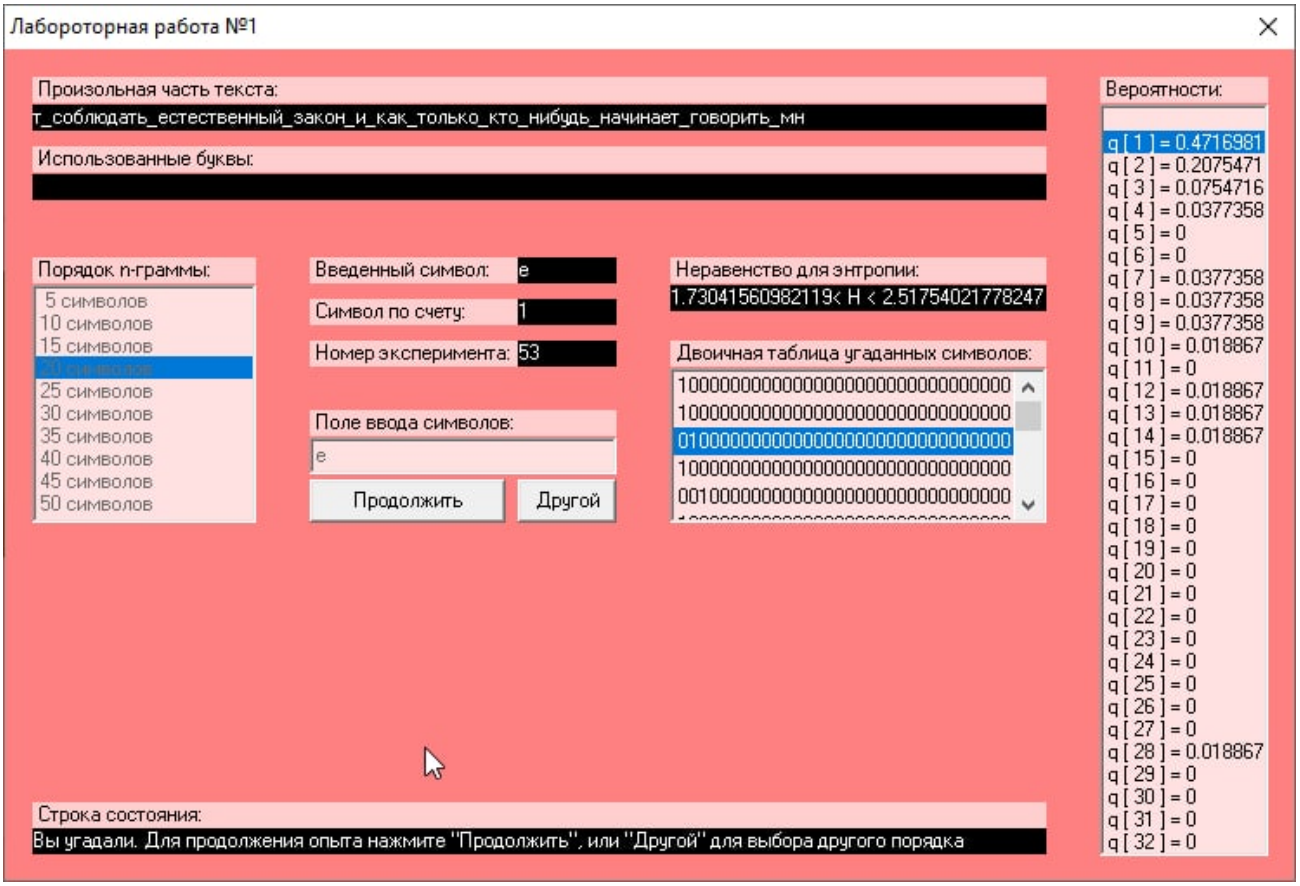
Частота біграмм з перетином (перші 10 найчастіші):

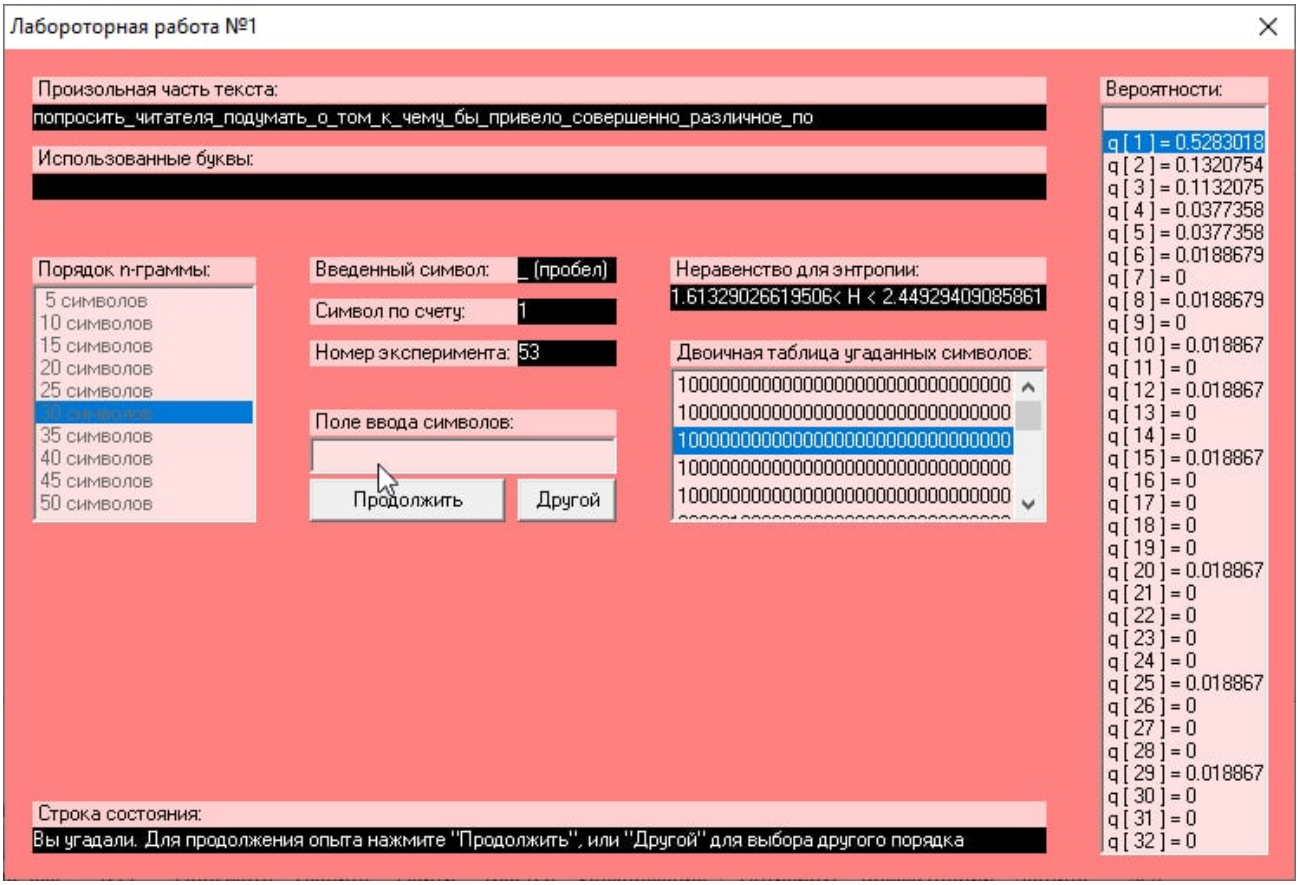
	Frequency
__	0,831717
о_	0,003393
е_	0,002975
_в	0,002858
_с	0,002633
и_	0,002574
а_	0,002535
_п	0,002521
то	0,002303
_н	0,002269

	Текст без пробілів	Текст з пробілами
H1	4.462	1.219
Redundant (H1)	9.917%	76.411%
H2	4.145204	0.996901
Redundant(H2)	16.329%	80.863%
H2 (with intersec.)	4.145589	0.996954
Redundant (H2 with intersec.)	16.321%	81.002%

2. За допомогою програми CoolPinkProgram оцінити значення H (10), H (20) , H (30)







$$2.156 < H(10) < 2.980$$

H(10)ср = 2.568

Надлишковість мови R для H(10):

$$R(max) = 1 - 2.156 / 5 = 0.569$$
$$R(min) = 1 - 2.980 / 5 = 0.404$$

0.404 < R < 0.569

R(ср) = **0.486**

$$R = 1 - \frac{H_\infty}{H_0}$$

$H_0 = \log_2 32 = 5$

$$1.730 < H(20) < 2.518$$

H(20)ср = 2.124

Надлишковість мови R для H(20):

$$R(max) = 1 - 1.730 / 5 = 0.654$$

$$R = 1 - \frac{H_\infty}{H_0}$$

$H_0 = \log_2 32 = 5$

R(min)= 1 - 2.518 / 5 = 0.496

0.496 < R < 0.654

R(ср) = **0.575**

1.613 < H(30) < 2.449

H(30)ср = 2.031

Надлишковість мови R для H(30):

R(max) = 1 - 1.613 / 5 = 0.677
R(min) = 1 - 2.449 / 5 = 0.510

0.510 < R < 0.677

R(ср) = **0.594**

$$R = 1 - \frac{H_\infty}{H_0}$$

$H_0 = \log_2 32 = 5$

Середнє значення надлишковісті:

R = (0.486 + 0.575 + 0.594) / 3 = **0.552**

Висновок: виконуючи дану лабораторну роботу ми набули практичних навичок щодо оцінки ентропії, у нас з'явлося справжнє уявлення про ентропію, про надлишковість мови.

При використанні власноруч написаної программи були отримані данні в яких можна помітити велику різницю між данними від двох різних видів обробки тексту. Дана різниця пов'язана із тим, що текст має дуже велику кількість пустих рядків із пробілами.

За допомогою програми CoolPinkProgram оцінити значення H (10), H (20) , H (30). Для кождого знайшли надлишковість, в наших експериментах R середній вийшов 0.552 (~55%). Але це природня надлишковість, як і всіх інших мовах.